DOCUMENT RESUME

ED 319 768 TM 014 962

AUTHOR Westfall, Philip Jean-Louis; D'Costa, Ayres G.
TITLE Improving Prediction by Correcting Test Scores for Person Disturbances Using the Rasch Model.
PUB DATE 87
NOTE 5p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classification; College Entrance Examinations; *Error of Measurement; French; Goodness of Fit; Item Response Theory; *Language Tests; Postsecondary Education; *Predictive Validity; *Scoring; Student Placement; Test Validity; *True Scores
IDENTIFIERS Air Force Academy CO; *Person Fit Measures; Placement Tests; *Rasch Model

ABSTRACT

This study, based on the Rasch model, used R. M. Smith's (1986) classification of measurement disturbances to assess the Rasch model approach to error control and statistical prediction. Partitioning the error component into a person component, an item-person interaction component, and a random unexplained error component has the net effect of reducing the error variance and improving statistical prediction and classification efficiency. The specific goal of this study is to determine the possibility of using the Item and Person Analysis Rasch Model (IPARM) to correct a person's score to eliminate person disturbances. The United States Air Force Placement and Validation Examination was administered to 1,200 freshmen entering the Air Force Academy in Colorado Springs; the French language portion of the test was used. Test data were analyzed using IPARM to identify individual response disturbances and to estimate each person's true score. IPARM can assess item disturbances as well as person disturbances, but this study focused on the latter. Essentially, IPARM uses a person's total score to establish statistical expectancies for each item the individual attempted. Test result correlations with French grades and class rankings indicated that the IPARM-based corrections are useful for improvement of prediction and classification. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED319768

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Ayres G. D'Costa

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IMPROVING PREDICTION BY CORRECTING TEST SCORES FOR PERSON DISTURBANCES USING THE RASCH MODEL

Philip Jean-Louis Westfall and Ayres G. D'Costa
U.S. Air Force Academy The Ohio State University

TM014962

BEST COPY AVAILABLE

Improving prediction by correcting test scores for person disturbances using the Rasch Model

Philip Jean-Louis Westfall and Ayres G. D'Costa
U.S. Air Force Academy The Ohio State University

Background

Several efforts have been made to refine the classical analysis of an observed score as the sum of a true score component for the individual and a random error component. Pike (1978) identified four components: a true score, a primary test-specific component, a secondary test-specific component and the usual random error component. This study is based upon the Rasch Model (Wright and Stone, 1979) and utilized Smith's (1986) classification of measurement disturbances into two types: those associated with the person, and those related to the item-person interaction.

The focus of this study is on Smith's approach to error control, rather than on Pike's efforts at true score analysis. Our interest is statistical prediction or classification, rather than an analysis of the sources of intellectual ability. We will ignore the dimensionality of the true score and assume it to be unidimensional. Partitioning the error component into a person component, an item-person interaction component, and a random unexplained error component, has the net effect of reducing the error variance and improving statistical prediction/ classification efficiency.

Smith identified eight person-related disturbances: start-up test anxiety, excessive cautiousness or plodding, copying/ cheating, external distractions, illness, systematic guessing, random guessing, and excessive carelessness or sloppiness. He also identified five disturbances due to item-person interaction: guessing when item is too difficult, over-confidence, item over- or under-learned, item type-style bias, and item information bias.

The research question of this study is as follows: What if it were possible using the Rasch model, specifically the IPARM (Item and Person Analysis using the Rasch Model) program, to correct a person's score to eliminate these disturbances? Would the resulting score, presumably a better representation of the individual's ability, improve prediction of future performance? IPARM has been developed by Smith, Wright and Green (1987).

Study Design and Rationale

The USAF Placement and Validation Examination (PLAVAL) was administered to the entire group of 1200 freshmen entering the

TM 0149162

Academy. Cadets are classified into three levels of additional French language training (low, intermediate, advanced) based on PLAVAL scores. Misclassification has serious cost implications because of financial, personal, and opportunistic effects.

PLAVAL data were analyzed using IPARM so as to identify individual response disturbances and estimate a person's true score. IPARM can look at item disturbances as well as person disturbances, but this study focussed only on the latter.

Essentially, IPARM uses a person's total score to establish statistical expectancies for each item the individual attempted. If a low-ability person correctly answered a very difficult question, the standardized residual (very much like the Sato Caution Index) becomes large. Whereas the Sato Caution Index (Harnish, 1983) is reported as is, IPARM notes that these residuals approximate a chi-square distribution and can provide a test of goodness of fit. Fit statistics can be provided for one item, a group of items, or an entire test. It is the analyses of groups of items that lead to discovering the nature of the misfit when present, and applying an appropriate correction. Test items can be divided into subgroups representing difficulty, order of presentation, item styles, or content. Each subgroup is analyzed as a test independent of other subgroups, with fit statistics characterizing an individual's expected response pattern. Estimates of ability are then determined for each subgroup. Items contributing most to an individual's misfit statistics are eliminated.

The research design compared the success of the classification of the 1200 USAF cadets into the three French class levels using three methods of prediction: based on the usual raw-score, a regular Rasch-model-based true score, and an IPARM-based corrected true score. First, success was measured in terms of correlation with French grades earned by each cadet in the freshman year. Second, success was measured in terms of the French class level classified into.

Results:

Table 1 presents the correlation coefficients (Pearson) for the three methods relative to the three French classes.

	Low	Inter	Advanced
Raw Score	0.48	0.46	0.44
Rasch Only	0.52	0.51	0.45
IPARM	0.63	0.63	0.48

Table 2 below presents the percent correctly classified, using discriminant analysis techniques, for the three French classes.

	Low	Inter	Advanced
Raw Score	56.9	52.4	51.6
Rasch Only	61.5	52.4	45.3
IPARM	63.1	66.7	53.1

Both sets of analysis indicate the superiority of the IPARM technique over using Rasch method by itself, and clearly over the usual Raw Score method. The advanced group appears to be the most hazardous to predict to. This is understandable given statistical regression.

Implications

This study indicates the utility of using the IPARM-based corrections at least for the improvement of prediction and classification. It appears that person-related disturbances can impinge upon the estimate of true score provided by tests, and that this error in turn can adversely affect prediction and classification. The use of IPARM techniques has educational utility and needs further exploration as a diagnostic/remediation device.

Our study did not explore whether the disturbances we corrected for were real, in the sense that the individual was personally aware of them. We were unable to conduct individual verification of the disturbances identified by IPARM. This is a serious limitation of this study and one that we had no control over because of restrictions imposed by USAF policy. We recommend that IPARM be used to provide individual feedback and assistance with remediation of test-taking "disturbances."

References

Harnish, D. (1983). "Item Response Patterns: Applications for Educational Practice." Journal of Educational Measurement, 20, 2, 191-206.

Pike, L. W. (1978). Short-term instruction, test-wiseness, and the Scholastic Aptitude Test. A literature review with research recommendations (ETS Research Report 78-2). Princeton, NJ: Educational Testing Service.

Smith, R. M. (1986). Diagnosing and correcting measurement disturbances. Unpublished manuscript. The Ohio State University, Columbus, OH.

Smith, R. M., Wright, B. D. & Green, K. (1987). Applications of the Rasch Model. Unpublished manuscript.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA Press.